

Discriminant Dimensionality Reduction Improves Alzheimer's And Parkinson's Disease Classification In Raman Spectroscopy Data

Sharon Sunny*, Bharath C Boyapati*, Mahesan Niranjan | School of Electronics and Computer Science * Contributed equally to the work
Christopher Kipps | Dept. Neurology, Wessex Neurological Centre, University Hospital Southampton
Sumeet Mahajan | School of Chemistry & Chemical Engineering

KEY TAKE AWAY MESSAGES

- * An efficient dimensionality reduction technique boosts the predictive performance of a machine learning classifier while reducing computational overhead
- * Discriminative features generated by utilising class label information better distinguish the samples in different classes
- * An optimal preprocessing strategy is critical as the downstream tasks are sensitive to noise in the data
- * Non-negative matrix factorisation and its variants offer physical interpretability of how much each feature contributes to the generated features

INTRODUCTION

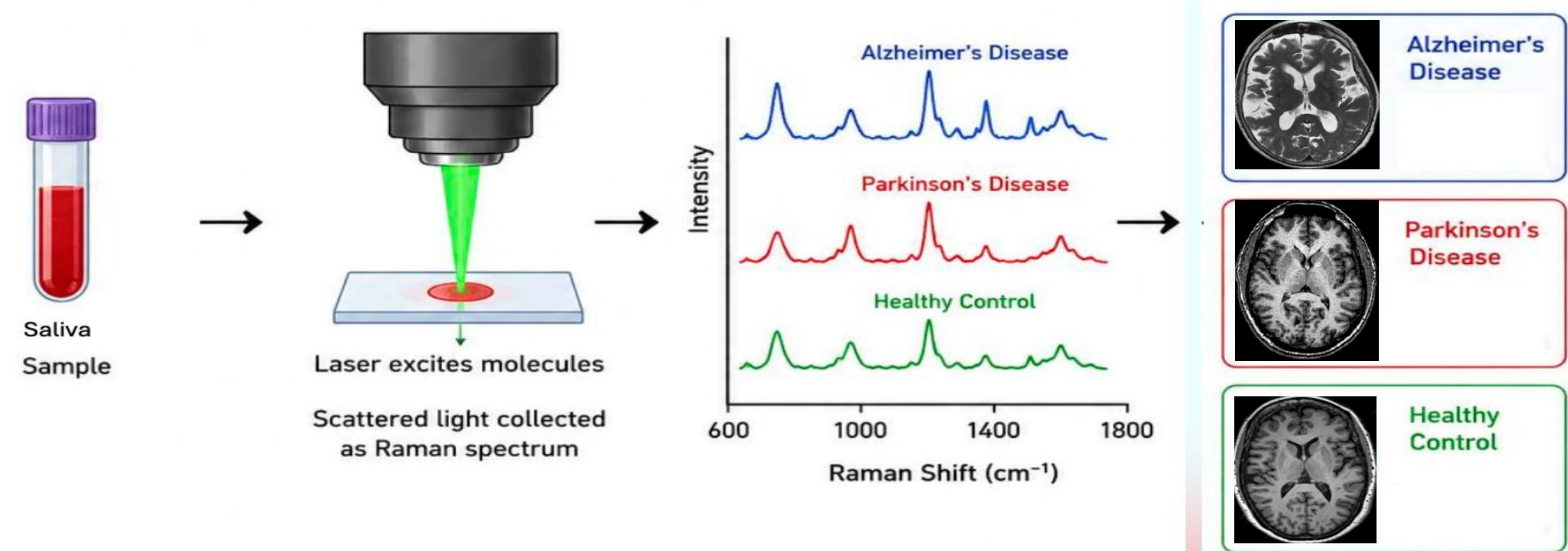
Why Raman spectroscopy?

- Offers label free, non-destructive molecular fingerprinting of samples
- Characteristic vibrational signatures of proteins, lipids, and nucleic acids provide a biochemical readout of disease state from biological fluids
- Non-invasive option for disease diagnosis

CHALLENGE

Big p small n [1]

- * Small patient cohort
- * Raman spectrum contains intensity values recorded at hundreds of different wavelengths
- * Machine learning models overfit
- * Applied feature reduction methods are purely based on variance or data reconstruction
- * Variance based methods lack physical interpretability



OUR SOLUTION

Discriminative Non-negative Matrix Factorisation [2]

- * **Generate class discriminant features:** incorporates class label information into the NMF objective, pushing basis vectors to better separate classes
- * **Offers physical interpretability:** non-negativity constraints mean learned basis components correspond to additive, physically meaningful sub-patterns
- * **Guarantees sparsity of features:** ensures each sample is represented by only a few active basis components, improving feature compactness and reducing overfitting

EXISTING METHODS

PCA $\max U^T C U$
subject to $U^T U = I$

NMF $\min \|X - UV\|_F$
subject to $U_{ij}, V_{ij} > 0$

MATERIALS & METHOD

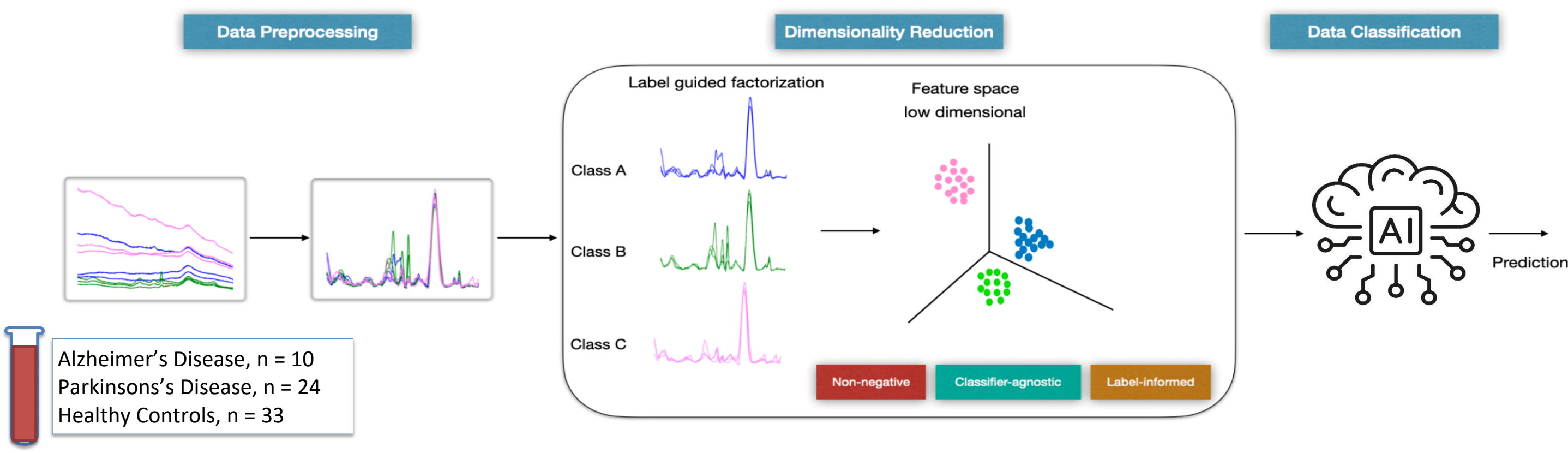
DNMF

Objective function:

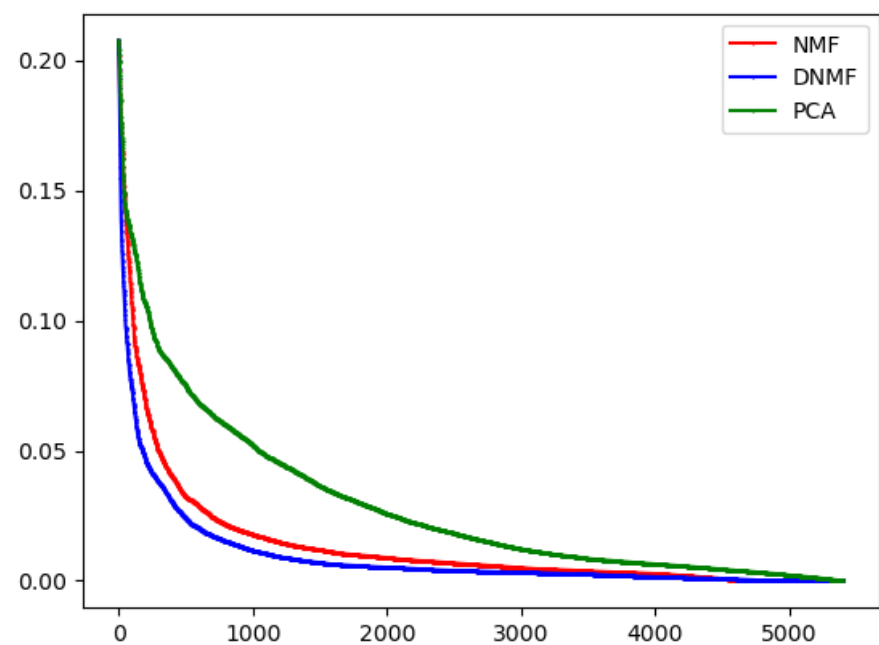
$$\min_{U,V \geq 0} \|X - UV\|^2 + \alpha \|Q - AV^T\|^2$$

Multiplicative update rules:

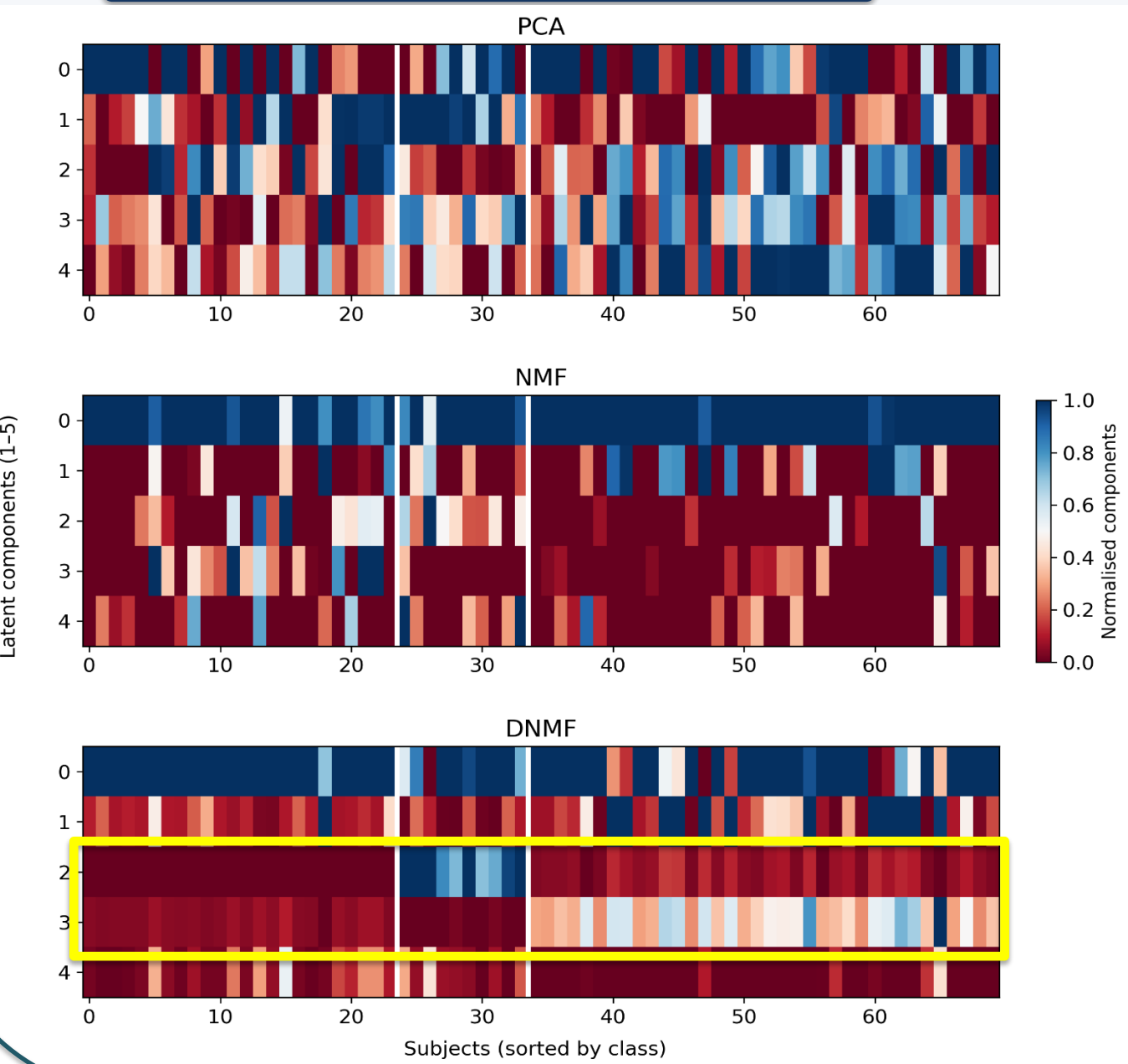
$$u_{ik} = u_{ik} \frac{(XV)_{ik}}{(UV^T V)_{ik}}$$
$$v_{jk} = v_{jk} \frac{(X^T U + \alpha(VA^T A)^- + \alpha(Q^T A)^+)_{jk}}{(VU^T U + \alpha(VA^T A)^+ + \alpha(Q^T A)^-)_{jk}}$$
$$A = QV(V^T V)^{-1}$$



SIGNIFICANT FEATURES



GENERATED COMPONENTS



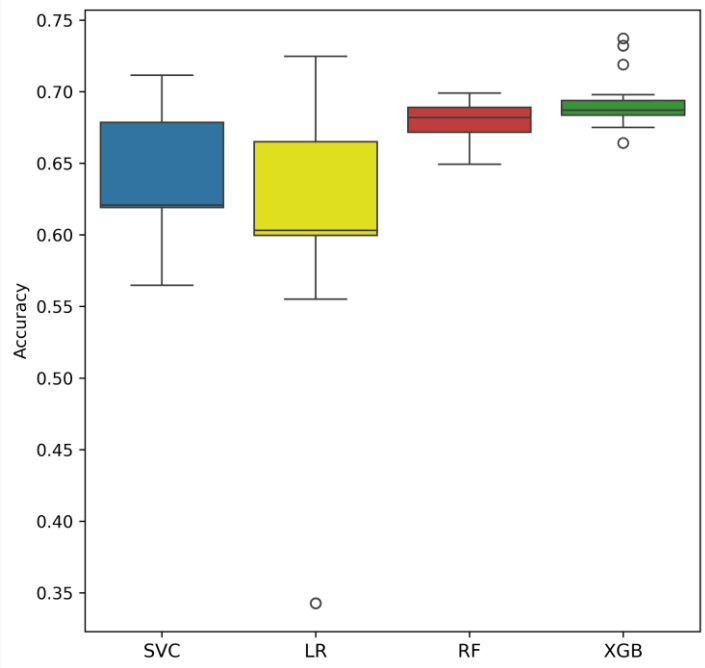
RESULTS

Dataset	DNMF	Vanilla NMF	PCA	Top-10	Top-60%	AllFeatures
PD-AD	0.86±0.08	0.84±0.12	0.84±0.11	0.75±0.06	0.80±0.07	0.66±0.07
Covid-19	0.88±0.06	0.86±0.06	0.87±0.08	0.69±0.09	0.85±0.06	0.83±0.05
Colorectal cancer	0.99±0.00	0.97±0.01	0.99±0.00	0.80±0.04	0.96±0.01	0.97±0.01
Ovarian cancer	0.91±0.05	0.75±0.13	0.83±0.06	0.49±0.28	0.37±0.13	0.37±0.13
Head & neck cancer						
Saliva dry	0.84±0.11	0.81±0.09	0.79±0.10	0.65±0.15	0.77±0.17	0.79±0.13
Saliva native	0.76±0.10	0.68±0.09	0.57±0.08	0.58±0.13	0.69±0.19	0.53±0.28
Plasma native	0.85±0.23	0.72±0.17	0.76±0.20	0.74±0.18	0.73±0.22	0.74±0.22
Plasma dry	0.74±0.15	0.77±0.14	0.78±0.13	0.64±0.13	0.73±0.18	0.74±0.17
Endometriosis ovarian cancer	0.62±0.02	0.63±0.02	0.62±0.02	0.61±0.03	0.59±0.02	0.58±0.03

DISCUSSION

- Inclusion of all features impose adverse effect on classification
- Systematic feature generation improves classifier performance
- Carefully designed pre-processing pipeline cleans the noisy raw data, enabling better classification
- Parameter selection and optimisation has significant effect on decision making capability of a model

Variation in accuracy values across different preprocessing pipelines



FUTURE WORK

- * Probabilistic model to facilitate uncertainty of prediction [3]
- * Minimum description length to select subspace size [4]
- * Integrate geometry of the datapoints in addition to the class information

REFERENCES

1. Johnstone, I. M., & Titterton, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1906), 4237.
2. Babaei, M., Tsoukalas, S., Babaei, M., Rigoll, G., & Datcu, M. (2016). Discriminative nonnegative matrix factorization for dimensionality reduction. *Neurocomputing*, 173, 212-223.
3. Squires, S., Prügel-Bennett, A., & Niranjan, M. (2025). A variational autoencoder for probabilistic on-negative matrix factorisation. *Pattern Analysis and Applications*, 28(4), 1-10.
4. Squires, S., Prügel-Bennett, A., & Niranjan, M. (2017). Rank selection in nonnegative matrix factorization using minimum description length. *Neural computation*, 29(8), 2164-2176.
5. Fan, K., Cai, X., & Niranjan, M. (2024). Non-negative subspace feature representation for few-shot learning in medical imaging. *Image and Vision Computing*, 152, 105334.
6. Liu, J., Fan, K., Cai, X., & Niranjan, M. (2024). Few-shot learning for inference in medical imaging with subspace feature representations. *PLoS one*, 19(11), e0309368.

CONTACT

